Statistical properties of global precipitation in the NCEP GFS model and TMPA observations for data assimilation

Guo-Yuan Lien[1,2], Eugenia Kalnay[1], Takemasa Miyoshi[1,2,3], and George J. Huffman[4]

[1]Department of Atmospheric and Oceanic Science, University of Maryland, College Park, Maryland, USA

[2]RIKEN Advanced Institute for Computational Science, Kobe, Japan

[3]Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan

[4]Mesoscale Atmospheric Processes Laboratory, NASA Goddard Space Flight Center, Greenbelt, Maryland

Submitted to Monthly Weather Review

(Revised on August 12, 2015)

Corresponding author's address: Guo-Yuan Lien, Data Assimilation Research Team, RIKEN Advanced Institute for Computational Science, 7-1-26, Minatojima-minami-machi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

E-mail: Guo-Yuan Lien, guo-yuan.lien@riken.jp
Eugenia Kalnay, ekalnay@atmos.umd.edu
Takemasa Miyoshi, takemasa.miyoshi@riken.jp
George J. Huffman, george.j.huffman@nasa.gov

Abstract

There are many issues regarding the assimilation of satellite precipitation data into numerical models, including the non-Gaussian error distributions associated with precipitation, and large model and observation errors. As a result, it is not easy to improve the model forecast beyond a few hours by assimilating precipitation. To identify the challenges and propose practical solutions to assimilation of precipitation, statistics are calculated for global precipitation in a low-resolution NCEP Global Forecasting System (GFS) model and the TRMM Multisatellite Precipitation Analysis (TMPA). The samples are constructed using the same model with the same forecast period, observation variables, and resolution as planned in the follow-on GFS/TMPA precipitation assimilation experiments presented in the companion paper.

The statistical results indicate that the T62 and T126 GFS models generally have positive bias in precipitation compared to the TMPA observations, and that the simulation of the marine stratocumulus precipitation is problematic in the T62 GFS model. It is necessary to apply to precipitation either the commonly used logarithm transformation or the newly proposed Gaussian transformation to obtain a better relationship between the model and observational precipitation. When the Gaussian transformations are separately applied to the model and observational precipitation, they serve as a bias correction that corrects the amplitude-dependent biases. In addition, using a spatially and/or temporally averaged precipitation variable, such as the 6-hour accumulated precipitation, should be advantageous for precipitation assimilation.

Key words: data assimilation, precipitation, Gaussian anamorphosis, bias correction.

1. Introduction

In recent years, several global precipitation estimations from a variety of remote sensing platforms have become available, such as the Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis (TMPA; Huffman et al. 2007, 2010) and the Global Satellite Mapping of Precipitation (GSMaP; Ushio et al. 2009). Meanwhile, many efforts to assimilate precipitation observations have also been made (e.g., Tsuyuki 1996, 1997; Falkovich et al. 2000; Davolio and Buzzi 2004; Koizumi et al. 2005; Mesinger et al. 2006). However, serious difficulties still remain in assimilating the precipitation data. For example, most of data assimilation schemes, including the variational methods and the ensemble Kalman filter (EnKF) methods, assume Gaussian error distributions for both observations and model backgrounds. If the error distribution is not Gaussian, the analysis may not be optimal. Since the precipitation-related variables are far from Gaussian, the non-Gaussianity issue becomes a severe problem for precipitation assimilation. Besides, both the model errors and observation errors are important issues for precipitation assimilation. As a consequence, a widely shared experience is that the precipitation assimilation can be useful in improving the model analyses, but the forecast improvement is usually limited to the first few forecast hours (e.g., Falkovich et al. 2000; Davolio and Buzzi 2004; Tsuyuki and Miyoshi 2007). These issues have been discussed and summarized in several articles, such as Errico et al. (2007), Bauer et al. (2011), and Lien et al. (2013; LKM2013 hereafter). Notwithstanding these difficulties, several recent studies have shown some usefulness of precipitation assimilation (Lopez 2011, 2013; Zupanski et al. 2011; Zhang et al. 2013).

A variable transformation technique is a computationally feasible solution to mitigate the non-Gaussianity issue in realistic geophysical data assimilation systems (Bocquet et al. 2010;

Amezcua and van Leeuwen 2014). For precipitation data assimilation, the precipitation values are usually transformed by a logarithmic function before assimilating them into the model (e.g., Lopez 2011). Instead of the logarithmic transformation, LKM2013 proposed to apply the Gaussian anamorphosis method to precipitation based on its model climatology, under the assumption that a forecast variable with more Gaussian climatological distribution would result in a more Gaussian error distribution. With this transformation, they succeeded in showing effective assimilation of global precipitation in their proof-of-concept observing system simulation experiments (OSSEs), using a simplified general circulation model and the local ensemble transform Kalman filter (LETKF). In their experiments, precipitation assimilation not only improves the analyses but also improves the model forecasts over the entire 5-day forecast period in their experiments.

Although a significant forecast improvement by precipitation assimilation was demonstrated in LKM2013 with an idealized system, in real systems improvements are generally very limited or even absent. The distinct challenges associated with the use of realistic model and real observations include the large and unknown errors related not only to the moist physical parameterization in the model but also to the observations. Since both the model precipitation and the observations could have large different types of errors, the long-term statistics of these two quantities may be very different, which is harmful to the data assimilation use. Therefore, before performing real precipitation data assimilation, it is worthwhile to first investigate the statistical characteristics of precipitation in both model and observation datasets which we would like to use, presented in this paper.

We investigate the differences in probability distributions between the precipitation in a series of short-term model forecasts and a precipitation observation dataset, to isolate the

different characteristics of the real model and observations. It is noted that the challenges introduced by these differences could not be addressed in LKM2013 since they used the identical-twin OSSE method. Here we use more realistic settings: the National Centers for Environmental Prediction (NCEP) Global Forecasting System (GFS), run at a low-resolution, and the TMPA data as the precipitation observations. Given the low resolution feasible in our study, the main focus of our work is assimilation of the global large-scale precipitation, which could be particularly important for improving medium-range model forecasts. Since the probability distributions are dependent on the use (or lack of use) of variable transformations, the results with different transformation methods will be investigated. We also show the correlation between model forecasts and observations at each grid point in a map. Several suggestions for real-data precipitation assimilation are made in the concluding section of this article. Although we choose to use the NCEP GFS model and the TMPA data to study the precipitation data assimilation, the same analysis can also be performed with other models and observation datasets.

The paper is organized as follows. The GFS model and TMPA observations are briefly introduced in Section 2. Section 3 describes the transformation methods we will use in the precipitation statistics. A series of statistical results are then presented in the following sections: Section 4 shows the cumulative distribution functions (CDFs) of the precipitation data, which will be used to define the Gaussian transformation of precipitation; Section 5 shows the joint probability distribution diagrams between the model precipitation and precipitation observations and compares the results in terms of the transformation methods, the temporal integration of precipitation, and the resolution of precipitation data. Section 6 presents the geographic distribution of correlation scores between these two variables. Concluding remarks and suggestions for the precipitation assimilation are given in Section 7. In addition, the successful

88 assimilation of the TMPA data following the guidance derived from this study will be presented
89 in a separate paper (Lien et al. 2015b; LMK2015b hereafter).

2. The model and observations

90 The GFS model is the operational global NWP model used at the NCEP. It is one of the
91 major world state-of-the-art operational NWP models and provides main model guidance for
92 weather forecasting in the United States. The GFS model can be run at various spectral
93 resolutions on a hybrid sigma/pressure coordinate. In this study we focus on the large-scale
94 global precipitation and also consider the computational constraints, so the experiments and
95 analyses are done with two lower-resolution configurations: T62 and T126 (roughly equivalent
96 to 200 km and 100 km horizontal resolutions) with 64 vertical levels (L64). The convection and
97 precipitation are parameterized using a modified simplified-Arakawa-Schubert (SAS) scheme
98 (Pan and Wu 1995; Han and Pan 2011), considering both deep and shallow convection.

99 The TRMM Multi-satellite Precipitation Analysis (Huffman et al. 2007, 2010) is a gridded
100 precipitation dataset compiled from multiple satellite sensors. It has a global coverage from 50°S
101 to 50°N with 0.25° spatial resolution and 3-hour temporal resolution. The variable provided by
102 the TMPA is the estimated surface precipitation rate. The primary data sources are the low-earth-
103 orbit (LEO) satellites such as the Microwave Imager (TMI) on TRMM, the Special Sensor
104 Microwave Imager (SSMI) and Special Sensor Microwave Imager/Sounder (SSMIS) on the
105 Defense Meteorological Satellite Program (DMSP) satellites, the Advanced Micro-wave
106 Scanning Radiometer-Earth Observing System (AMSR-E) on Aqua, the Advanced Microwave
107 Sounding Unit-B (AMSU-B) on the National Oceanic and Atmospheric Administration (NOAA)
108 satellite series, and the Microwave Humidity Sounder (MHS) on both the NOAA and the
109 EUMETSAT MetOp series. The microwave satellite observations have a strong physical

relationship to the hydrometeors and thus the surface precipitation, but they are spatially and temporally inhomogeneous. To fill the gaps left from the LEO sensors, the infrared (IR) data collected by the geosynchronous-earth-orbit (GEO) satellites are used as the secondary data sources with calibration by the microwave precipitation estimates, though the accuracy of precipitation derived from the IR is lower. For the research version (i.e., not in real time) of the TMPA, these satellite-derived precipitation amounts are further rescaled based on several monthly rain gauge analyses to achieve accurate statistics in the climatological scale, while in the real-time version the satellite-derived precipitation is rescaled with a climatological correction to the research version. With the above data processing procedure, the TMPA has very high (> 95%) data coverage rate (Figure 1a), thus becoming a potential good observational source for the assimilation of global precipitation. In this study, we use the version 7 of the TMPA research products, labeled as 3B42, released in 2012 (Huffman et al. 2012). The climatological mean daily precipitation computed from the 14-year TMPA data (1998–2011) is shown in Figure 1b.

To make the 0.25°-resolution TMPA data correspond to the lower resolutions of the T62/T126 GFS model, we pre-process the precipitation rate data, upscaling the original TMPA grids to the T62 or T126 Gaussian grids used by the GFS model using an area-conserving remapping.

3. Transformation of Precipitation

In this section, several transformations for precipitation assimilation are described, including the widely used logarithm transformation, and the transformation based on Gaussian anamorphosis used in previous studies such as Simon and Bertino (2009), Schöniger et al. (2012), and LKM2013. The transformations have a profound impact on the statistical results shown in later sections.

a. Logarithm transformation

The logarithm transformation

$$\tilde{y} = \ln(y + \alpha) \tag{1}$$

is a simple and frequently used way to transform precipitation. Here, y is the original variable, $\tilde{y}$ is the transformed variable, and α is a tunable constant added to prevent the singularity at zero precipitation ($y = 0$). Using the logarithm transformation, Lopez (2011) successfully assimilated the NCEP stage IV precipitation analysis over the eastern United States, and Lopez (2013) presented experimental results of assimilation of the 6-hourly accumulated precipitation observations measured by the rain gauges at synoptic stations.

b. Gaussian transformation

The logarithm transformation may be helpful for precipitation assimilation in some regions, seasons, or precipitation types, but a globally invariant analytical transformation may not be applicable to every case. Therefore, following LKM2013, we will also examine the effect of the Gaussian transformation on the precipitation statistics. Here we briefly summarize the formulation of the Gaussian transformation in LKM2013 and explain the changes made in this study after LKM2013.

1) General formula

The transformations is made by equating the two CDFs of the original variable (y) and the transformed variable ($\tilde{y}$):

$$\tilde{F}(\tilde{y}) = F(y) \text{ , or} \tag{2}$$

$$\tilde{y} = \tilde{F}^{-1}[F(y)] \text{ ,} \tag{3}$$

147 where F is the CDF of y, $\tilde{F}$ is the CDF of $\tilde{y}$, and $\tilde{F}^{-1}$ is the inverse function of $\tilde{F}$. By definition,
148 the CDFs are bounded within $[0, 1]$. The CDF of the original variable (F) is empirically
149 determined from samples, and the CDF of the transformed variable ($\tilde{F}$) can be arbitrarily chosen
150 so that the transformed variable can have any desired distribution. If we choose

$$\tilde{F}(\tilde{y}) = F^G(\tilde{y}) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\tilde{y}}{\sqrt{2}}\right)\right], \tag{4}$$

151 which is the CDF of a standard normal distribution with zero mean and unit variance, and erf is
152 the error function, then

$$F^{G^{-1}}(P) = \sqrt{2}\,\mathrm{erf}^{-1}(2P - 1) \tag{5}$$

153 where P is the cumulative probability, so that it becomes a "Gaussian anamorphosis"
154 (Wackernagel 2003):

$$\tilde{y} = F^{G^{-1}}[F(y)]\,. \tag{6}$$

155 In this way, the transformed variable ($\tilde{y}$) becomes a Gaussian variable. The use of the Gaussian
156 anamorphosis has appeared in several geophysical data assimilation studies (e.g., Simon and
157 Bertino 2009, 2012; Schöniger et al. 2012). We call this method "Gaussian transformation"
158 hereafter.

159 Figure 2 provides an illustration the Gaussian transformation procedure. It displays the 10-
160 year climatological probability density function (PDF) and CDF of the original and transformed
161 precipitation in both the GFS model forecasts and the TMPA dataset, at three selected locations
162 for the 11–20 January period. The collection of the model and observational precipitation
163 samples will be discussed in later sections, but here we use the plots to visualize the method. The
164 transformation starts from Figure 2a (e, i), which are the very non-Gaussian PDFs of the original

165 variables. The red color stands for the model precipitation and the green color stands for the
166 observational precipitation. Their CDFs are then calculated [Figure 2c (g, k)]. Using the inverse
167 CDF of the standard normal distribution $F^{G^{-1}}$, the cumulative probability values are converted
168 into the transformed variables $\tilde{y}$, whose CDFs shown in Figure 2d (h, l) and PDFs in Figure 2b
169 (f, j). It is important to note that the precipitation distribution contains a great portion of zero
170 values, shown as a delta function in the PDFs and a discontinuity in the CDFs, which need to be
171 treated in a special manner. Following LKM2013, all the zero values are represented by half of
172 the zero precipitation cumulative probability (i.e., the median; solid circles in Figure 2) during
173 the transformation:

$$F(0) = \frac{1}{2} P_c \,. \tag{7}$$

174 where P_c is the zero precipitation probability in the climatology. In this way, the zero
175 precipitation is still a delta function in the transformed variable, but it is located at a certain
176 distance away from the trace precipitation values.

177 This method transforms the climatological distribution of the model forecast variable into a
178 Gaussian distribution, but this does not necessarily make the *background error distributions*
179 Gaussian, as required in the EnKF data assimilation (e.g., Ott et al. 2004). However, it is
180 reasonable to assume that a forecast variable with more Gaussian climatological distribution
181 would result in more Gaussian error distribution (LKM2013). It is difficult to validate this
182 assumption using the climatological data in this study but we do provide a validation of this
183 assumption in the follow-on paper (LMK2015b) using the actual experimental data from the
184 cycling LETKF data assimilation.

185 It is worth mentioning that this CDF-based transformation of precipitation has also been
186 used in some climate studies, though they are not related to data assimilation. For example, the
187 Standardized Precipitation Index (SPI) (McKee et al. 1993; Guttman 1999) commonly used to
188 study drought is defined based on a similar method, but the time scales of precipitation
189 accumulations they have focused on are much longer than the 6 hours used in weather data
190 assimilation.

2) Computation of the CDFs and transformations

191 Some technical details are described in this subsection. First, we regard all precipitation
192 values smaller than 0.06 mm $(6h)^{-1}$ as "zero precipitation" because small values in the model or
193 observational precipitation data would be not meaningful. This value is close to the threshold
194 used in LKM2013, 0.1 mm $(6h)^{-1}$.

195 Second, extreme values with cumulative distribution less than 0.001 and greater than 0.999
196 are set to 0.001 and 0.999, respectively. Consequently, when the original values fall outside the
197 range in the climatological samples, they will be transformed to -3.09 and 3.09. It is noted for
198 reference that Simon and Bertino (2012) also discussed this problem and they used parametric
199 linear tails to form their transformation.

200 Third, we derive the CDFs from precipitation samples using constant-width bins with
201 respect to the cumulative probability in [0, 1], not with respect to the precipitation amount as it
202 might be intuitively done. Two hundred bins are used. The CDFs are thus represented by the 201
203 (including 0 and 1) discretized precipitation amounts at each cumulative distribution levels at a
204 0.005 increment. When we need to compute $F(y)$ for a given precipitation value y, we perform a
205 linear interpolation from the two nearby data points. Compared to binning with respect to the

precipitation amount, this method can more precisely represent the CDF curves using the same number of the bins, particularly for large precipitation values.

3) Separate Gaussian transformation applied to model background and observations

Following the methods described above, we can apply the Gaussian transformation to the GFS model and the TMPA data. However, there is an important difference between the Gaussian transformation used in LKM2013 and in this study. In LKM2013, the transformation was defined purely based on the 10-year model precipitation climatology, and so the same transformation was used for both the model precipitation and the observed precipitation. There was no need to consider the transformations of the model precipitation and the observed precipitation separately because the work used an identical-twin configuration so that the two CDFs are identical. In contrast, in this study with a realistic model and real observations, the transformations need to be defined separately for model precipitation and observations (see red and green colors in Figure 2). Specifically, the transformation of the model precipitation is performed based on the CDF computed from the model climatology; and the transformation of the precipitation observations is performed based on the CDF computed from the observation climatology. In this way, the model climatology and the observation climatology are first converted to the same 0–1 scale of their cumulative distribution using the corresponding transformation (Figure 2d), then the same $F^{G^{-1}}$ is applied to obtain the Gaussian variables (Figure 2b). Therefore, this method can essentially remove the climatological bias between these two variables that is dependent on the precipitation values, referred to as the "amplitude-dependent bias". The effect of the separate transformations can be large because the precipitation distribution of the model and observational precipitation can be very different at some regions (e.g., Figure 2i–l), which will be discussed in later sections.

4. Cumulative distribution functions of the climatological precipitation data

We first construct the empirical CDFs for both the GFS model background precipitation and the TMPA observations, based on their climatological samples. These model and observational CDFs will be compared, and they will also be used in defining the Gaussian precipitation transformation. For a relevant comparison useful for guiding the assimilation of precipitation, we examine the quantities that are used in the data assimilation, which depend on the design of any specific data assimilation system. We now describe how we collect the 10-year samples of the model background precipitation and observations in correspondence with our proposed 4D-LETKF experiments.

Figure 3 shows a schematic of the sample preparation. First, for the model precipitation, we would like to have the "background values" which are usually the short-term (e.g., 6 hours) forecasts from the analyses. In our system of 4D-LETKF, forecast variables within the period from 3 to 9 hours will be used as the model background (Hunt et al. 2004; Miyoshi and Yamane 2007). Therefore, we conduct a series of 9-hour GFS model forecasts at desired resolutions (T62 and T126 in this study) every 6 hours initialized from 10-year (2001–2010) CFSR reanalysis data, then the 3 to 9 hour forecasts are collected to form a series of model backgrounds. The GFS model outputs forecast fields every hour in the form of the instantaneous precipitation rate, thus we can either pick up the precipitation rates every 3 hours corresponding to the TMPA observations or compute the 6-hour accumulated precipitation centered at time t by

$$\mathrm{P(6h)}_t = \frac{1}{2}\mathrm{Pr}_{t-3} + \sum_{t\prime=t-2}^{t+2} \mathrm{Pr}_{t\prime} + \frac{1}{2}\mathrm{Pr}_{t+3}\,, \tag{8}$$

where Pr_t is the precipitation rate (mm h^{-1}) at time t. Note that although we could directly use reanalysis precipitation as the model precipitation samples without performing the short-term

248 forecasts, doing in the manner of this study should be preferable because the existing reanalysis
249 dataset may be produced in a way that is different from our proposed data assimilation system
250 (e.g., different configurations of the forecast model), and the specific variable used in the data
251 assimilation, such as the accumulated precipitation within the 3–9 hour forecast may be not
252 provided in the reanalysis dataset.

253 For the observations, the same 10-year (2001–2010) data should be collected to form a
254 series of equivalent observational data. The original TMPA data are provided with the 3-hourly
255 precipitation rate at a 0.25° longitude-latitude resolution. After upscaling the TMPA data to the
256 Gaussian grids used by the T62/T126 GFS model, either the instantaneous precipitation rate as in
257 its original form, or the 6-hour accumulated precipitation amount can be used to compute the
258 statistics. The 6-hour accumulated precipitation centered at time t is computed by a weighted
259 average

$$\mathrm{P(6h)}_t = \frac{3}{2}\mathrm{Pr}_{t-3} + 3\ \mathrm{Pr}_t + \frac{3}{2}\mathrm{Pr}_{t+3}\ . \tag{9}$$

260

261 After collecting large samples of model background and observational precipitation values,
262 their CDFs are computed using the method described in Section 3.b, for each T62 grid point and
263 each 10-day period of year (3 periods per month; 36 periods in total); i.e.,

$$F = F(y; \text{location, period of year})\ , \tag{10}$$

264 where y can be either model or observed 6-hour accumulated precipitation in their original value,
265 and F is the CDF, as previously defined in Equations (2) and (3). The real data contain large
266 spatial and temporal variabilities. Therefore, to create a more "continuous" CDF field smoothly
267 varying in space and time, we include all data within 500-km radius and ± 2 periods (± 20 days)

when computing the CDF at each grid point and each period. This choice also increases the sample sizes and thus reduces the sampling errors. The grid numbers within the 500-km radius are about 20 for the T62 resolution and 80 for the T126 resolution (changing with the geographical location), so the total grid numbers used to construct the CDF for each point are roughly $10\ (\text{year}) \times 365\ (\text{day/year}) \times 4\ (\text{cycle/day}) \times (5\ \text{period}/36\ \text{period}) \times \{20, 80\} \cong \{4 \times 10^4, 1.6 \times 10^5\}$ for the {T62, T126} resolution, respectively.

We already presented in Figure 2 the examples of CDFs at 3 different types of regions in the extratropics (Maryland), in the tropics, and in the marine stratocumulus region for demonstrating how to construct the Gaussian transformation. The marine stratocumulus region shows a large discrepancy between the CDFs of the model and observational precipitation. To visualize the entire CDF field as a function of the geographic location, we plot the maps of precipitation amounts at various cumulative distribution levels also for the period of 11–20 January for both the TMPA data and the T62 GFS model backgrounds (Figure 4). By comparing the fields at the same cumulative distribution levels, it is clearly found that the model has a positive bias compared to the observations since the amounts in Figure 4b, d, f are generally greater than those in Figure 4a, c, e. Positive biases are generally seen in the other seasons (not shown). In terms of geographical patterns, the CDF fields of the model and observations agree reasonably well in most regions. However, in some particular regions, they actually have a large disagreement. For example, the GFS forecast shows a local maximum in the precipitation amount at both 30% and 60% cumulative distribution levels (Figure 4b, d) in the Pacific Ocean west to the Southern America (at about 20°S), but this local maximum does not appear in the TMPA data (Figure 4a, c, e). This is the region corresponding to the marine stratocumulus precipitation.

290 This discrepancy in these regions is most apparent in maps showing the probability of zero
291 precipitation. As shown in Figure 5, the most significant differences in the zero precipitation
292 probability between the model and observations are found over the regions where the marine
293 stratocumulus are formed over cold waters, including the subtropical eastern Pacific in both
294 northern and southern hemispheres (west of North and South America), and west of Australia
295 and Africa. In the TMPA data, it rarely rains in these regions (typically with 90% probability of
296 zero precipitation or 10% probability of nonzero precipitation; green open circle in Figure 2k, l),
297 but the model drizzle is too frequent, with typically 80% probability of nonzero precipitation (red
298 open circle in Figure 2k, l). Several studies of the marine stratocumulus (vanZanten et al. 2005;
299 Leon et al. 2008) indicate that the real nonzero precipitation probability is not as high as the
300 model climatology here, favoring the TMPA data. The precipitation parameterization in the low
301 resolution T62 GFS model may be unable to correctly simulate the low level of marine
302 stratocumulus precipitation. However, Huffman (2007) documented that the TMPA also has a
303 low precipitation bias over ocean due to lack of sensitivity of microwave imager to light
304 precipitation, so these large differences could come from both high bias in the model and low
305 bias in the TMPA data. Since in this paper we do not attempt to improve either the model or the
306 observations, a reasonable strategy is to not to assimilate the precipitation data in regions where
307 the disagreement between the model background and the observations is large.

5. Joint probability distributions

308 In this section we use the joint probability distribution diagrams to more clearly show the
309 relationship between the model background precipitation and the precipitation observations. All
310 data points in the 10-year samples are included in the statistics. Results with different

311 transformation methods, different variables (i.e., precipitation rate vs. accumulated
312 precipitation), and different resolutions will be shown and discussed.

a. Original data vs. logarithm transformed precipitation

313 Figure 6 shows the joint probability distribution diagrams between the 6-hour accumulated
314 precipitation in the T62 GFS model background and in the TMPA data upscaled to the same T62
315 grids. Different transformation methods are used in each subplot. Only nonzero precipitation is
316 shown in the figures because when the zero precipitation is also plotted, it just adds two saturated
317 lines along the x-axis $(\tilde{y}, \tilde{y}_{\text{zero}})$ and y-axis $(\tilde{y}_{\text{zero}}, \tilde{y})$ representing the abundance of zero
318 precipitation in either the model background or the observation data (not shown). One would
319 expect that the maximum probability regions should be located along the one-to-one diagonal
320 line for a variable that is useful for data assimilation. However, when the joint probability
321 distribution diagram is plotted without a transformation method (Figure 6a), we barely see any
322 correlation in precipitation between the model background and the observations[1]. The probability
323 of small precipitation amounts is saturated and not oriented along the one-to-one line. This partly
324 explains why the original precipitation is not a good variable for data assimilation and an
325 appropriate transformation of precipitation is needed.

326 When we calculate the joint probability using logarithm transformed precipitation [without
327 adding a constant in the logarithmic function; $\alpha = 0$ in Equation (1)] (Figure 6b), the curved line
328 of the maximum probability (indicated with a red dashed curve) is clearly seen. This maximum
329 probability curve is to the right of the one-to-one line, indicating an amplitude-dependent
330 positive bias of the model precipitation when compared to the TMPA data. In this data

[1] In this case, the R^2 value computed from linear regression shown in the figure may not be particularly meaningful, since the correlation largely comes from the off-diagonal regions.

assimilation study, we do not argue whether the model precipitation or the TMPA data is more correct, but it is clearly better to remove this bias before data assimilation. For example, bias correction schemes have been widely used in the modern satellite radiance data assimilation (e.g., Derber and Wu 1998; Dee 2005).

In addition, an interesting fact is found when the "modified" logarithm is used [i.e., a constant $\alpha = 0.6$ mm $(6h)^{-1}$ is added in the transformation; Equation (1)]. In Figure 6c, saturation in the small precipitation amounts, as in Figure 6a, is seen again. The maximum probability curve near the one-to-one line is still retained but it is less obvious than in Figure 6b. Therefore, from this joint probability distribution diagram, it is inferred that the use of a too large constant α in the logarithm transformation may not be a good solution, since it makes the behavior of the transformed variable in the small precipitation amounts similar to the original variable, and thus reduces the discrimination for small amounts. A careful choice of the α value is thus essential.

b. Precipitation rate vs. accumulated precipitation

Figure 7a shows the same diagrams but for the instantaneous precipitation rate ($\alpha = 0$ in the logarithm transformation). Comparing with Figure 6b, it is clear that the correlation with the precipitation rate is worse than that with the accumulated precipitation amount. In particular, a multimodal feature is seen in the model precipitation. The precipitation rate produced from the T62 GFS model tends to be concentrated at several ranges (roughly [-3, -2], [-1.5, -1], and [0, 1] in the logarithm-transformed value), which could be related to some deficiencies of the precipitation parameterization at this low resolution. The lower correlation may also be a result of the timing error of the precipitation parameterization scheme. The instantaneous precipitation rate is too sensitive to the timing error, which is common for the precipitation produced from cumulus parameterizations. For example, Chao (2013) showed that cumulus precipitation

353 schemes can have large systematic errors in the precipitation diurnal cycle over the land.
354 Therefore, although the accumulation of precipitation discards the information of the time
355 variations of the precipitation within the 6-hour assimilation window, the 6-hour accumulated
356 value of precipitation would be still a better variable than the precipitation rate when used in data
357 assimilation. The successful assimilation of precipitation demonstrated by Lopez (2011, 2013)
358 also used the 6-hour accumulated precipitation. Nevertheless, we note that the model resolution
359 we use is fairly coarse, and the precipitation parameterization could perform better in a higher
360 resolution model.

c. Resolution (T62 vs. T126)

361 The same diagram of Figure 6b but based on the higher resolution results (6-hour
362 accumulated precipitation) is shown in Figure 7b. We carry out all the same processes used in
363 Figure 3 at the T126 resolution. At this resolution, the bias between the model and observational
364 precipitation is clearly smaller than that at the T62 resolution as seen in the joint probability
365 distribution diagrams (i.e., the deviation of the maximum probability line from the one-to-one
366 line in Figure 7b is smaller than that in Figure 6b); however, the correlation between the model
367 and observations also becomes slightly lower than that at T62 (i.e., 0.1625 vs. 0.1822 in R^2).
368 This is probably due to the larger random error in the higher resolution model and observation
369 data. By spatially averaging the field, this random error can be reduced (Huffman et al. 2010),
370 which may be easier for the precipitation assimilation.

371 However, there is certainly loss of information caused by upscaling the observation data to
372 lower resolution, and also a reduction in the accuracy of numerical models by using the low
373 resolution configuration. Therefore, the choice of the resolution may depend on the specific
374 purpose of the work. In this study, we propose that, for the purpose of improving large-scale

375 medium range forecasts, using the spatially averaged (i.e., upscaled) TMPA data would be a
376 reasonable choice. Indeed, we show in the companion paper (LMK2015b) that the assimilation
377 of the global large-scale (lower-resolution) precipitation field at the T62 resolution is able to
378 improve the 5-day model forecasts. We do not argue that the higher-resolution model or
379 observations are useless in precipitation assimilation, but that there is a "trade-off" between the
380 resolution and errors. Since it has been shown that model resolution leads to a large impact on
381 the precipitation forecasts (e.g., Wen et al. 2012), assimilating higher resolution precipitation
382 data and solving the issues regarding the random errors would be important research. Using a
383 higher resolution model that has better representation of precipitation processes but still
384 employing the spatial average in the observation operator could also be considered.

d. Gaussian transformed precipitation

385 Using the CDFs constructed in Section 4, we can define the Gaussian transformations of the
386 GFS model precipitation and the TMPA data following Section 3.b. Note again that the CDFs
387 are computed for each T62 grid point and each 10-day period of year, and smoothed by including
388 the nearby grids and times. Although this smoothing helps to construct a smooth CDF field and
389 thus a more continuous definition of the Gaussian transformation, the disadvantage of this
390 method is that the transformation would not be good in regions with intrinsically large gradient
391 of precipitation climatology, such as regions with complex terrain and orographic precipitation.

392 With the Gaussian transformation, the joint probability distribution diagrams are shown in
393 Figure 8. Figure 8a and d are the global results. Figure 8a uses the logarithm transformation
394 already shown before (Figure 6b), and Figure 8d is the same figure plotted with the Gaussian
395 transformed variables. The figure shows that with the Gaussian transformation, the distribution
396 of the precipitation variables become more normal, the maximum probability curve becomes

more collocated with the one-to-one line (i.e., the biases are reduced), and the correlation square (R^2) value increases slightly. In our transformation method defined for model and observations separately, the model climatology and the observation climatology are first converted to the same 0–1 scale (cumulative distribution), and then the same $F^{G^{-1}}$ is applied to obtain the Gaussian variables. Therefore, this method can effectively reduce the amplitude-dependent bias as seen in Figure 8a. We call this method a "CDF-based bias correction."

The same diagrams are then plotted with land data only (Figure 8b, e), ocean data only (Figure 8c, f), the northern hemisphere extratropics (20–50°N; Figure 9a, d), the tropical regions (20°N–20°S; Figure 9b, e), and the southern hemisphere extratropics (20–50°S; Figure 9c, f). Note that the TMPA only covers from 50°S to 50°N so the statistics are done within this extent. Overall, the improvements in the normality, centeredness, and correlations that we found in the global results are also found over the separate validation regions [except that the correlation slightly decreases over the ocean with the transformation (Figure 8c, f) but the change is small]. The amplitude-dependent biases are largely reduced in all regions. Using the logarithm transformation, the climatological distributions are skewed toward large precipitation amounts in the land and tropical regions where the convective precipitation is more prevalent, and toward small precipitation amounts in other regions. The skewness is less obvious in all regions when the Gaussian transformation is applied. As to the correlation, the increase of the correlation is particularly notable in the land region and in the northern hemisphere extratropics. In summary, we find that using separate Gaussian transformations applied to model background precipitation and observations, defined in terms of each grid point and each period of year, the climatological distributions of both these two variables are made more Gaussian, and their biases are significantly reduced.

6. Time correlation maps

Using the same 10-year samples of data, and the same Gaussian transformation, we also calculate the time correlations between the 6-hour accumulated model and observational precipitation at each grid point and each 10-day period of year so that their geographical distributions can be displayed. Similar to the CDF calculation, when computing the correlation at each grid point, the data within ± 2 periods (± 20 days) are considered together to obtain the temporally smoothed field. Thus this correlation score is a simple measure of the statistical "consistency" between the model and the observation climatologies. Figure 10 shows the global correlation maps in 4 different periods in January, April, July, and October. Overall, the dry area shows smaller correlations, which is expected because it may not easy to capture the small or infrequent precipitation amounts by the moist physical parameterization in the model. Besides, the correlation over ocean is generally much higher than that over land, except for the marine stratocumulus region, where the correlations are very low as shown from the discrepancy of the CDF statistics in Section 4. Over land, the desert areas (such as the Sahara) show persistent low correlations over the year probably because of the infrequent precipitation events and small precipitation values. The mountainous areas such as the Tibetan Plateau also show low correlations, which could be partly due to the problem of orographic precipitation in the satellite based estimates (Shige et al. 2012). Over the United States, the eastern area has higher correlation than the western area.

According to these time correlation maps, we think that the precipitation data distributed over the regions with reasonable correlations can be useful in the data assimilation to improve the model analyses and forecasts, but we hypothesize that the data over the too-small-correlation regions could be difficult to be used, possibly mainly because of the incapable precipitation

parameterization in the model. Therefore, it is motivated that we can set up some thresholds of the correlation values to reject the observations located over the small-correlation regions in the data assimilation process. We actually employed this idea in the real precipitation assimilation experiments (LMK2015b) and obtained a slight improvement than not using this criterion.

7. Concluding remarks and suggestions to precipitation assimilation

This article is the first part of our GFS/TMPA precipitation data assimilation study. In this part, we calculated statistics with the precipitation variable in the model background and observations from the point of view of data assimilation. To achieve meaningful statistics, the samples are carefully constructed using the same model with the same forecast period, observation variables, and resolution, as we planned to use in the real precipitation assimilation experiments (LMK2015b). These statistical results can indicate how to extract more useful information from the precipitation observations.

First of all, the errors of precipitation in numerical models can contribute to a substantial amount of the difficulties observed in the precipitation assimilation. For example, our statistical results indicate that the GFS model at both T62 and T126 resolution, generally has positive bias in precipitation as compared to the TMPA observations, and that it has a severe problem in parameterizing the marine stratocumulus precipitation. The "precipitation scale" is a key point of the problem. First, the method for creating precipitation in numerical models depends intrinsically on the different grid resolutions. When the grid resolution is low, the precipitation is mainly parameterized by cumulus convection schemes, but the behavior of the model precipitation varies with model resolution. For example, in the GFS model, precipitation at the T126 resolution is less biased than that at the T62 resolution, but the correlation to the observations is also slightly lower, presumably due to the increasing difficulty in collocating

forecasted and observed precipitation that comes with model resolution. When the grid resolution is sufficient to resolve convection, the microphysics parameterization schemes can take over the cumulus parameterization, and the behavior of the model precipitation may be very different (something not examined in this study). In addition, precipitation usually appears in random patches, especially for convective precipitation, leading to large random errors at high resolutions. The timing of the convective precipitation is also difficult to simulate by models. In addition, the high spatial and temporal variability further lead to large representativeness errors, which are also dependent upon resolution and important to data assimilation.

Performing spatial and/or temporal averages can effectively reduce these errors. Huffman et al. (2010) recommended TMPA users to create time/space averages that are appropriate to their application from the original fine-scale data. Bauer et al. (2011) also mentioned that using spatially/temporally smoothed precipitation data in assimilation can be beneficial. Based on similar arguments, accumulated precipitation (equivalent to a time average) is expected to be a better variable to be used in the data assimilation, rather than the instantaneous precipitation rate. However, this strategy may seem to contradict the continued pursuit of higher resolution, especially if we are able to afford high-resolution models and take high-resolution observations. We consider that this is a trade-off between resolution and errors. If the main goal is to improve the medium-range model forecasts, using a smoothed lower resolution precipitation to improve the large-scale analysis can be a reasonable choice. We note that the strategy needed for effective assimilation of convective scale precipitation such as meteorological radar observations could be quite different from the current context (e.g., Yussouf et al. 2013).

The ultimate solution to overcome the above issues would be attained by the improvement of the model precipitation parameterization and the satellite precipitation estimates. Strenuous

487 efforts have been made by the modeling (e.g., Han and Pan 2011) and remote sensing retrieval
488 communities (e.g., Tapiador et al. 2012). However, within the scope of our data assimilation
489 study, we do not attempt to improve the model or the observations. Our main goal is to optimally
490 use this imperfect observation dataset in this imperfect model, to improve the model forecasts of
491 both precipitation and non-precipitation variables, such as wind, temperature, and pressure, by
492 using appropriate error covariances in the data assimilation. To achieve this goal, we suggest
493 applying separate Gaussian transformations to model background and observational
494 precipitation, which can improve the Gaussianity of the variables while also effectively
495 removing the amplitude-dependent biases between these two variables. This idea is an extension
496 of the Gaussian precipitation transformation proposed for a perfect model by LKM2013 in which
497 the same transformation was applied to both model precipitation and observations.

498 However, since the transformation method is just an approximate way to mitigate the non-
499 Gaussianity issue in the data assimilation, and both the transformation and the bias correction are
500 constructed based only on the climatologies, there should be some limits of these transformation
501 and correction approaches. Therefore, precipitation observations that are deemed to be too bad to
502 be used may need to be rejected. Note that the statement "an observation is bad for assimilation"
503 is not necessarily because the observation itself is bad, but because the model is not capable of
504 making use of this observation in that location and time. The samples of the long-term model and
505 observational precipitation data we prepared in this study could be a useful reference to define
506 appropriate quality control criteria to assimilate only the "useful" precipitation observations.

507 Based on the discussion above, we suggest that the problems associated with the
508 assimilation of large-scale satellite precipitation data with the goal to improve the medium range
509 model forecasts should be addressed as follows:

- Non-Gaussianity of the precipitation variable: Apply the Gaussian transformation to both model and observational precipitation. In LKM2013, this was shown to be essential for effective assimilation of precipitation using the LETKF in the idealized experiments. LKM2013 also suggested performing the assimilation only when there are enough background members with nonzero precipitation.
- Inconsistent probability distributions of precipitation in model climatology and observation climatology: Define the Gaussian transformations for the model precipitation and the observational precipitation separately based on their own CDFs so that the amplitude-dependent bias is reduced. We call this method a "CDF-based bias correction."
- Timing errors of the precipitation: Use 6-h accumulated amounts.
- Deficient precipitation parameterization: Do not assimilate observations where the model is deficient. Appropriate quality control criteria (e.g., the climatological correlation scores between the model precipitation and observational precipitation) can be considered to keep only the precipitation observations that the model can effectively use.
- High-resolution observations contain large random errors: Perform spatial and/or temporal averages to reduce the random errors; upscale the observations to large-scale grids.

This guidance on the statistical approaches to precipitation assimilation were implemented and found to significantly improve the T62 5-day forecasts, shown in LMK2015b.

Acknowledgements

This study was mainly done as part of Guo-Yuan Lien's Ph.D. thesis work at the University of Maryland, partially supported by NASA grants NNX11AH39G, NNX11AL25G, NNX13AG68G, NOAA grants NA100OAR4310248, CICS-PAEK-LETKF11, and the Office of Naval Research (ONR) grant N000141010149 under the National Oceanographic Partnership Program (NOPP). We obtained a version of the GFS model from NOAA's Environmental Modeling Center (EMC) with the kind help of Henry Huang and Daryl Kleist, and the model was ported to our Linux cluster with the contribution by Tetsuro Miyachi. We also gratefully acknowledge the support from Japan Aerospace Exploration Agency (JAXA) Precipitation Measuring Mission (PMM).

References

Amezcua, J., and P. J. van Leeuwen, 2014: Gaussian anamorphosis in the analysis step of the EnKF: a joint state-variable/observation approach. *Tellus A*, **66**, doi:10.3402/tellusa.v66.23493.

Bauer, P., G. Ohring, C. Kummerow, and T. Auligne, 2011: Assimilating satellite observations of clouds and precipitation into NWP models. *Bull. Amer. Meteor. Soc.*, **92**, ES25–ES28, doi:10.1175/2011BAMS3182.1.

Bocquet, M., C. A. Pires, and L. Wu, 2010: Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.*, **138**, 2997–3023, doi:10.1175/2010MWR3164.1.

Chao, W. C., 2013: Catastrophe-concept-based cumulus parameterization: Correction of systematic errors in the precipitation diurnal cycle over land in a GCM. *J. Atmos. Sci.*, **70**, 3599–3614, doi:10.1175/JAS-D-13-022.1.

Davolio, S., and A. Buzzi, 2004: A nudging scheme for the assimilation of precipitation data into a mesoscale model. *Wea. Forecasting*, **19**, 855–871, doi:10.1175/1520-0434(2004)019<0855:ANSFTA>2.0.CO;2.

Dee, D. P., 2005: Bias and data assimilation. *Q.J.R. Meteorol. Soc.*, **131**, 3323–3343, doi:10.1256/qj.05.137.

Derber, J. C., and W.-S. Wu, 1998: The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon. Wea. Rev.*, **126**, 2287–2299, doi:10.1175/1520-0493(1998)126<2287:TUOTCC>2.0.CO;2.

Errico, R. M., P. Bauer, and J.-F. Mahfouf, 2007: Issues regarding the assimilation of cloud and precipitation data. *J. Atmos. Sci.*, **64**, 3785–3798, doi:10.1175/2006JAS2044.1.

Falkovich, A., E. Kalnay, S. Lord, and M. B. Mathur, 2000: A new method of observed rainfall assimilation in forecast models. *J. Appl. Meteorol.*, **39**, 1282–1298, doi:10.1175/1520-0450(2000)039<1282:ANMOOR>2.0.CO;2.

Guttman, N. B., 1999: Accepting the Standardized Precipitation Index: A calculation algorithm. *J. Am. Water. Resour. As.*, **35**, 311–322, doi:10.1111/j.1752-1688.1999.tb03592.x.

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, doi:10.1175/WAF-D-10-05038.1.

Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *J. Hydrometeor.*, **8**, 38–55, doi:10.1175/JHM560.1.

——, R. Adler, D. Bolvin, and E. Nelkin, 2010: The TRMM Multi-Satellite Precipitation Analysis (TMPA). *Satellite Rainfall Applications for Surface Hydrology*, M. Gebremichael and F. Hossain, Eds., Springer Netherlands, 3–22.

573 Hunt, B. R., and Coauthors, 2004: Four-dimensional ensemble Kalman filtering. *Tellus A*, **56**,
574 doi:10.3402/tellusa.v56i4.14424.
575 Koizumi, K., Y. Ishikawa, and T. Tsuyuki, 2005: Assimilation of precipitation data to the JMA
576 mesoscale model with a four-dimensional variational method and its impact on precipitation
577 forecasts. *SOLA*, **1**, 45–48, doi:10.2151/sola.2005-013.
578 Leon, D. C., Z. Wang, and D. Liu, 2008: Climatology of drizzle in marine boundary layer clouds
579 based on 1 year of data from CloudSat and Cloud-Aerosol Lidar and Infrared Pathfinder Satellite
580 Observations (CALIPSO). *J. Geophys. Res.*, **113**, D00A14, doi:10.1029/2008JD009835.
581 Lien, G.-Y., E. Kalnay, and T. Miyoshi, 2013: Effective assimilation of global precipitation:
582 Simulation experiments. *Tellus A*, **65**, 19915, doi:10.3402/tellusa.v65i0.19915.
583 Lopez, P., 2011: Direct 4D-Var assimilation of NCEP stage IV radar and gauge precipitation
584 data at ECMWF. *Mon. Wea. Rev.*, **139**, 2098–2116, doi:10.1175/2010MWR3565.1.
585 ——, 2013: Experimental 4D-Var assimilation of SYNOP rain gauge data at ECMWF. *Mon.*
586 *Wea. Rev.*, **141**, 1527–1544, doi:10.1175/MWR-D-12-00024.1.
587 McKee, T. B., N. J. Doesken, and J. Kleist, 1993: The relationship of drought frequency and
588 duration to time scales. *Proc. 8th Conference on Applied Climatology*, Boston, MA, American
589 Meteorological Society, 179–183.
590 Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor.*
591 *Soc.*, **87**, 343–360, doi:10.1175/BAMS-87-3-343.
592 Miyoshi, T., and S. Yamane, 2007: Local ensemble transform Kalman filtering with an AGCM
593 at a T159/L48 resolution. *Mon. Wea. Rev.*, **135**, 3841–3861, doi:10.1175/2007MWR1873.1.
594 Ott, E., and Coauthors, 2004: A local ensemble Kalman filter for atmospheric data assimilation.
595 *Tellus A*, **56**, doi:10.3402/tellusa.v56i5.14462.
596 Pan, H.-L., and W.-S. Wu, 1995: *Implementing a mass flux convective parameterization package*
597 *for the NMC medium-range forecast model*. NMC Office Note,
598 http://www.lib.ncep.noaa.gov/ncepofficenotes/files/01408A42.pdf.
599 Schöniger, A., W. Nowak, and H.-J. Hendricks Franssen, 2012: Parameter estimation by
600 ensemble Kalman filters with transformed data: Approach and application to hydraulic
601 tomography. *Water Resour. Res.*, **48**, W04502, doi:10.1029/2011WR010462.
602 Shige, S., S. Kida, H. Ashiwake, T. Kubota, and K. Aonashi, 2012: Improvement of TMI rain
603 retrievals in mountainous areas. *J. Appl. Meteor. Climatol.*, **52**, 242–254, doi:10.1175/JAMC-D-
604 12-074.1.
605 Simon, E., and L. Bertino, 2009: Application of the Gaussian anamorphosis to assimilation in a
606 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment.
607 *Ocean Sci.*, **5**, 495–510, doi:10.5194/os-5-495-2009.

608 Simon, E., and L. Bertino, 2012: Gaussian anamorphosis extension of the DEnKF for combined
609 state parameter estimation: Application to a 1D ocean ecosystem model. *J. Marine Syst.*, **89**, 1–
610 18, doi:10.1016/j.jmarsys.2011.07.007.
611 Tapiador, F. J., and Coauthors, 2012: Global precipitation measurement: Methods, datasets and
612 applications. *Atmospheric Research*, **104–105**, 70–97, doi:10.1016/j.atmosres.2011.10.021.
613 Tsuyuki, T., 1996: Variational data assimilation in the tropics using precipitation data. Part II:
614 3D model. *Mon. Wea. Rev.*, **124**, 2545–2561, doi:10.1175/1520-
615 0493(1996)124<2545:VDAITT>2.0.CO;2.
616 ——, 1997: Variational data assimilation in the tropics using precipitation data. Part III:
617 Assimilation of SSM/I precipitation rates. *Mon. Wea. Rev.*, **125**, 1447–1464, doi:10.1175/1520-
618 0493(1997)125<1447:VDAITT>2.0.CO;2.
619 ——, and T. Miyoshi, 2007: Recent progress of data assimilation methods in meteorology. *J.*
620 *Meteor. Soc. Japan*, **85B**, 331–361, doi:10.2151/jmsj.85B.331.
621 Ushio, T., and Coauthors, 2009: A Kalman filter approach to the Global Satellite Mapping of
622 Precipitation (GSMaP) from combined passive microwave and infrared radiometric data. *J.*
623 *Meteor. Soc. Japan*, **87A**, 137–151, doi:10.2151/jmsj.87A.137.
624 vanZanten, M. C., B. Stevens, G. Vali, and D. H. Lenschow, 2005: Observations of drizzle in
625 nocturnal marine stratocumulus. *J. Atmos. Sci.*, **62**, 88–106, doi:10.1175/JAS-3355.1.
626 Wackernagel, H., 2003: *Multivariate Geostatistics*. Springer, 408 pp.
627 Wen, M., S. Yang, A. Vintzileos, W. Higgins, and R. Zhang, 2012: Impacts of model resolutions
628 and initial conditions on predictions of the Asian summer monsoon by the NCEP Climate
629 Forecast System. *Wea. Forecasting*, **27**, 629–646, doi:10.1175/WAF-D-11-00128.1.
630 Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013: The
631 ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic
632 supercell storm using single- and double-Moment microphysics schemes. *Mon. Wea. Rev.*, **141**,
633 3388–3412, doi:10.1175/MWR-D-12-00237.1.
634 Zhang, S. Q., M. Zupanski, A. Y. Hou, X. Lin, and S. H. Cheung, 2013: Assimilation of
635 precipitation-affected radiances in a cloud-resolving WRF ensemble data assimilation system.
636 *Mon. Wea. Rev.*, **141**, 754–772, doi:10.1175/MWR-D-12-00055.1.
637 Zupanski, D., S. Q. Zhang, M. Zupanski, A. Y. Hou, and S. H. Cheung, 2011: A prototype WRF-
638 based ensemble data assimilation system for dynamically downscaling satellite precipitation
639 observations. *J. Hydrometeor.*, **12**, 118–134, doi:10.1175/2010JHM1271.1.
640

Figures

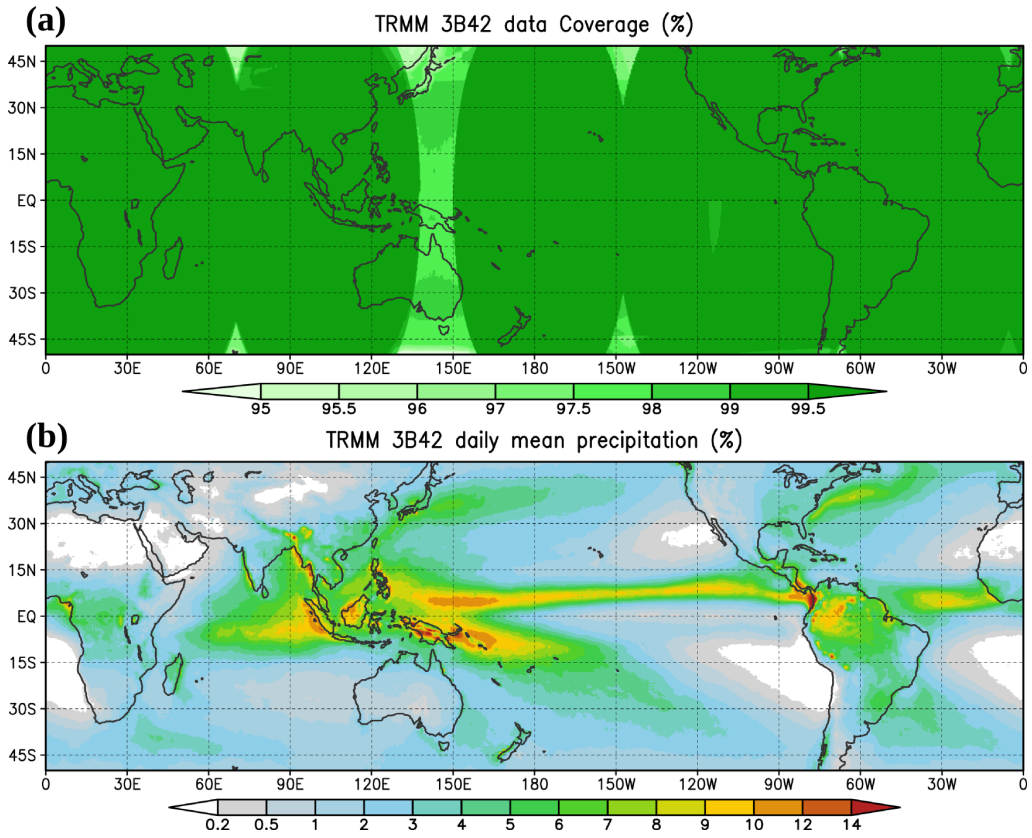

Figure 1: (a) The data coverage rate (%) and (b) the mean daily precipitation (mm) of the 14-year (1998-2011) TRMM Multi-satellite Precipitation Analysis. Note that the coverage in (a) is greater than 95% in most areas.

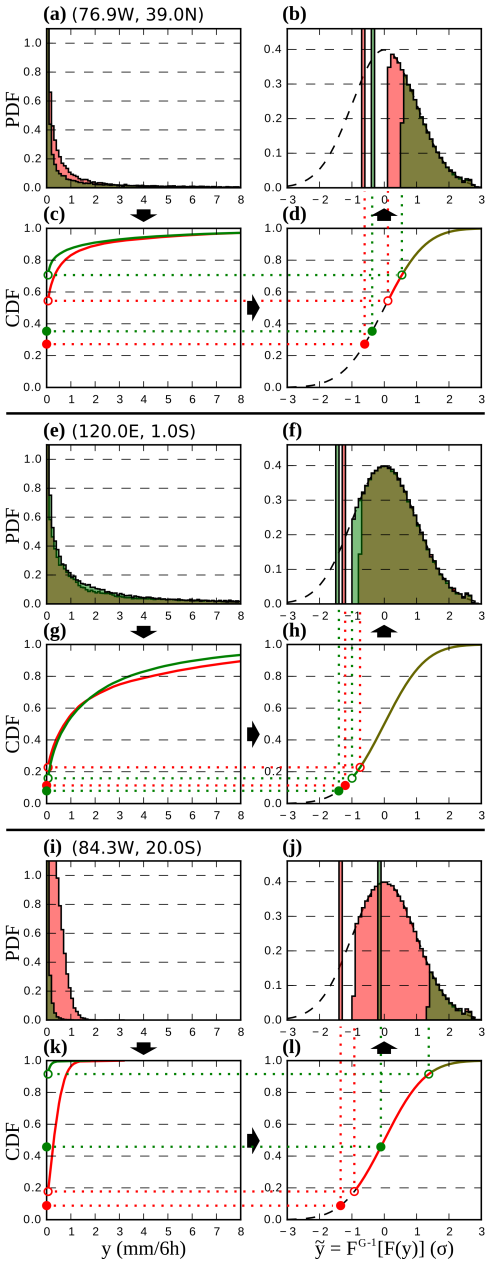

Figure 2: The probability density function and cumulative distribution function of the original precipitation and the transformed precipitation based on the 10-year model (red color) and observation (green color) climatologies. (a)–(d) A grid point in extratropics (76.9°W, 39.0°N); (e)–(h) A grid point in tropics (120.0°E, 1.0°S); (i)–(l) A grid point in a marine stratocumulus region west of South America (84.3°W, 20.0°S). All plots correspond to the 11–20 January period. The procedure of the Gaussian transformation is from (a) to (c), to (d), and to (b) as indicated by the arrows. The open circles correspond to the zero precipitation probability and the solid circles correspond to the half value (median) of the zero precipitation probability.

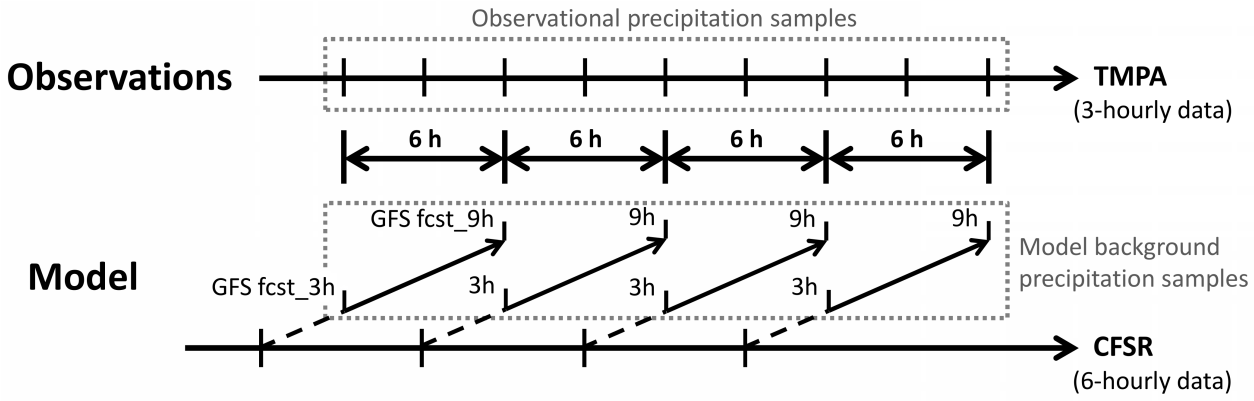

Figure 3: A schematic of the preparation of precipitation samples from the TMPA observation dataset and the GFS model forecasts. For precipitation observations, a 10-year series of the 3-hourly TMPA data is collected (top); for model background precipitation, equivalent 10-year data are formed from a series of 9-hour GFS model forecasts every 6 hours initialized from the 10-year CFSR reanalysis. In each forecast cycle, the forecast is conducted with the desired model configurtion and resolutions (T62 and T126 in this study), and only the 3 to 9 hour forecasts are used.

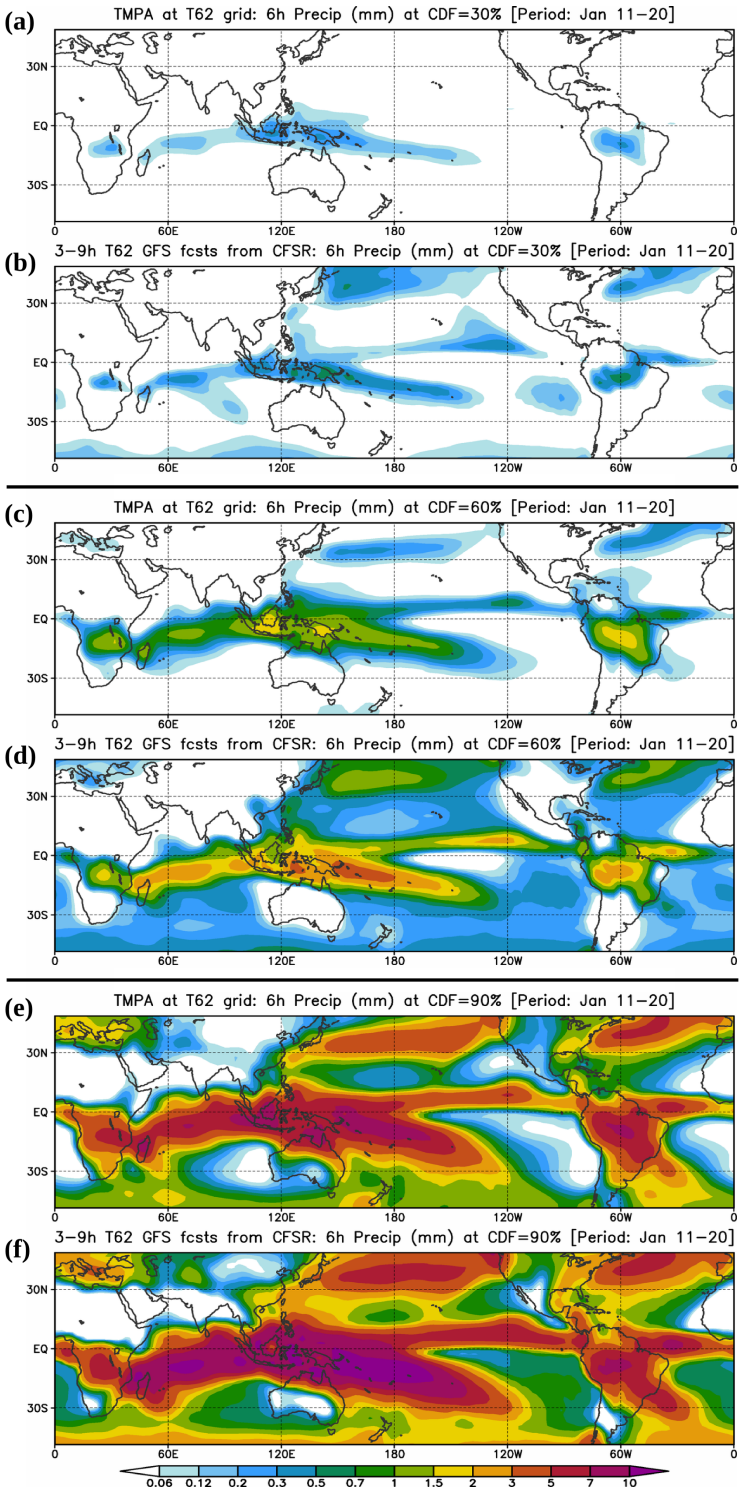

Figure 4: Comparison of TMPA and GFS precipitation amounts (mm) for different levels of the precipitation CDF. (a) (b) 30%, (c) (d) 60%, and (e) (f) 90% cumulative distribution levels during the 11–20 January period. (a) (c) (e) are TMPA data, and (b) (d) (f) are T62 GFS model forecasts.

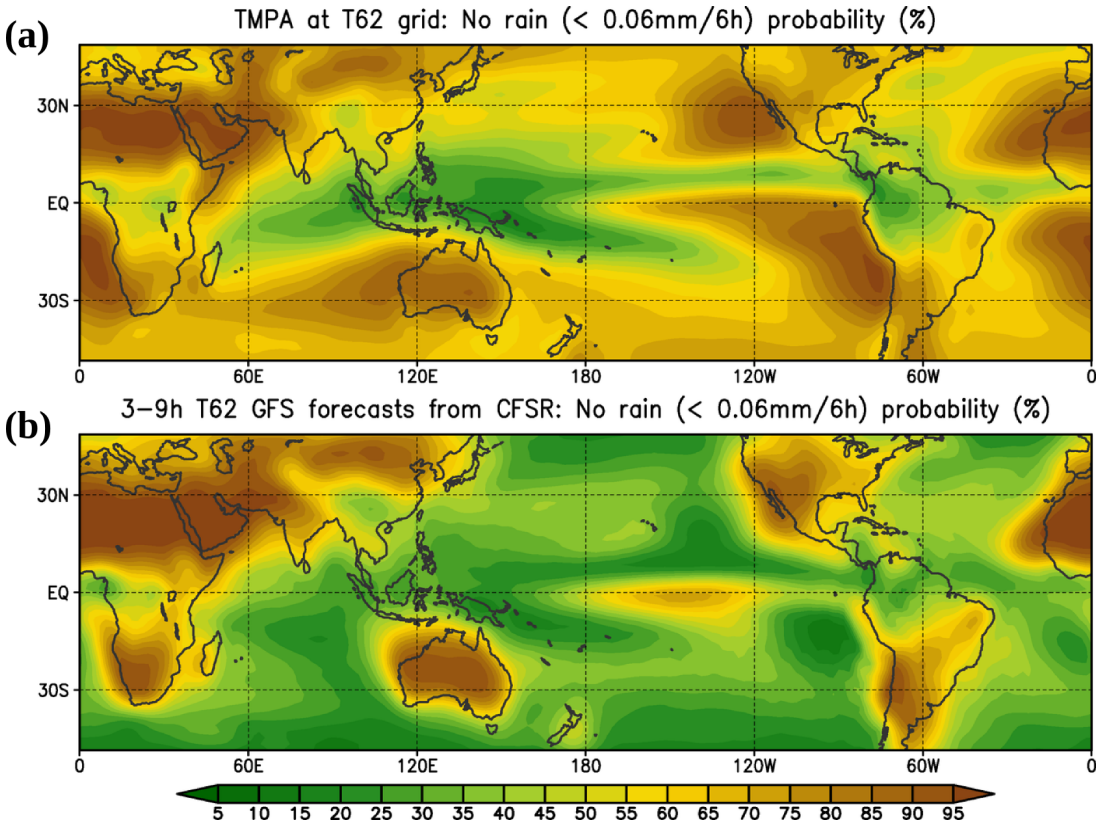

Figure 5: The maps of (all-season) zero precipitation probability (%) in (a) the TMPA data and (b) the T62 GFS model forecasts.

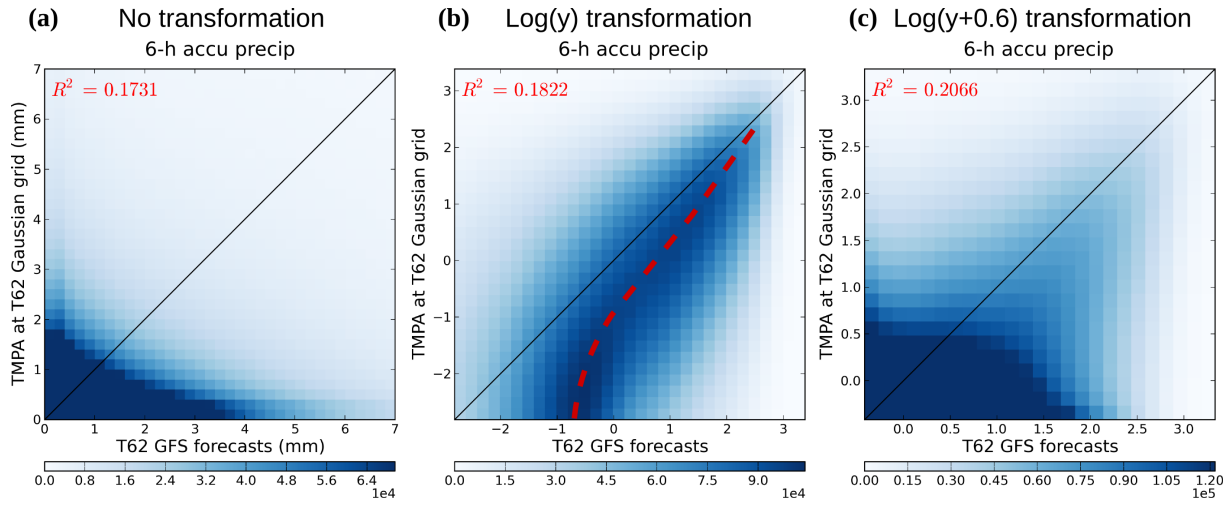

Figure 6: Joint probability distributions of the 6-hour accumulated precipitation with different transformation methods between the T62 GFS model background and the TMPA data upscaled to the same T62 grids. (a) No transformation (mm), (b) an exact logarithm transformation [$\alpha = 0$ in Equation (1)], (c) a "modified" logarithm transformation ($\alpha = 0.6$ mm) is applied to the precipitation variables. Only positive precipitation is shown.

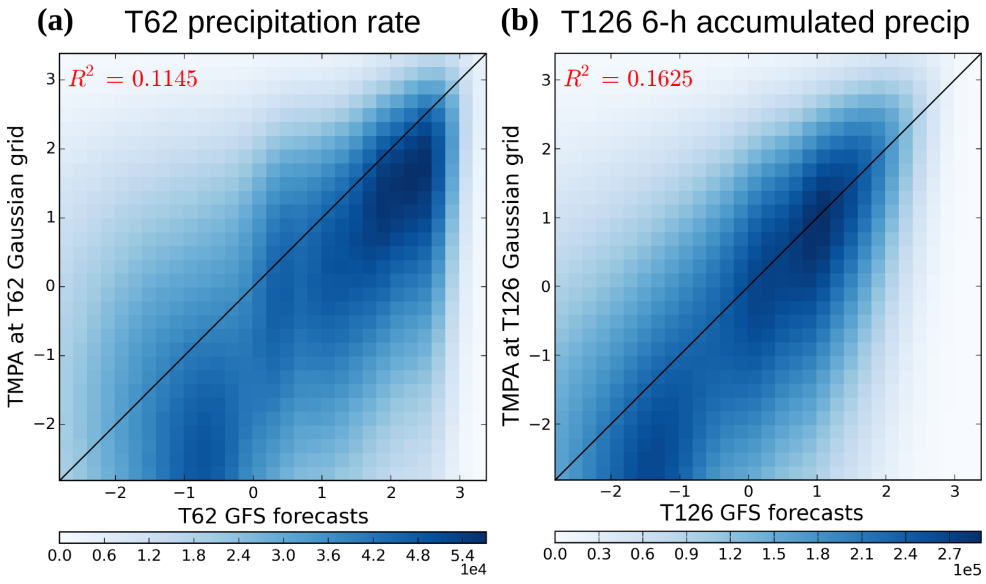

Figure 7: As Figure 6b, but for the logarithm-transformed (a) instantaneous precipitation rate [mm $(6h)^{-1}$ before the transformation] at the T62 resolution and (b) 6-hour accumulated precipitation (mm before the transformation) at the T126 resolution in both the GFS model background and the TMPA data.

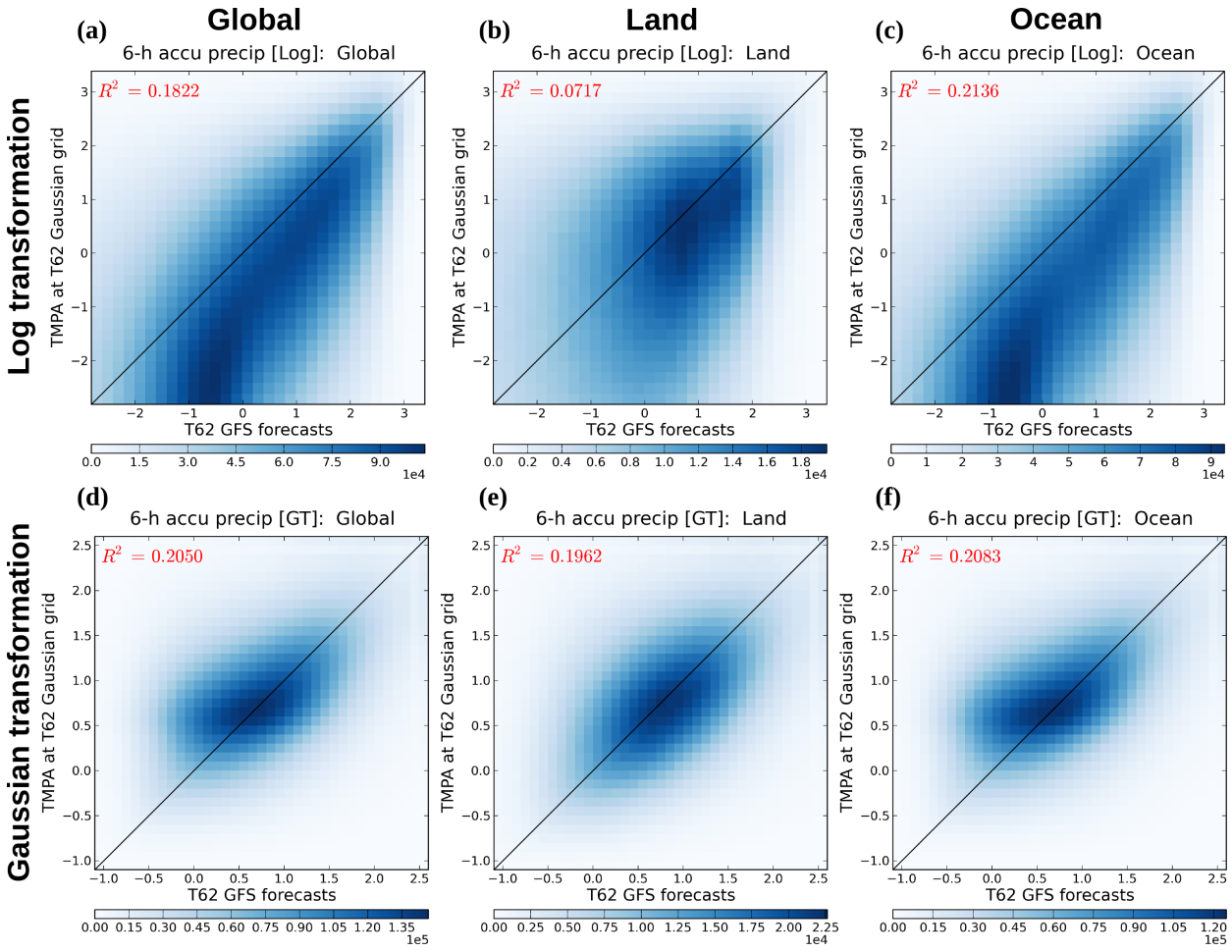

Figure 8: The joint probability distribution of (a)–(c) the logarithm-transformed ($\alpha = 0$) and (d)-(f) the Gaussian-transformed 6-hour accumulated precipitation between the T62 GFS model background and the TMPA data upscaled to the same T62 grids. (a) (d) Global results; (b) (e) only the precipitation over the land; (c) (f) only the precipitation over the ocean. Only positive precipitation is shown.

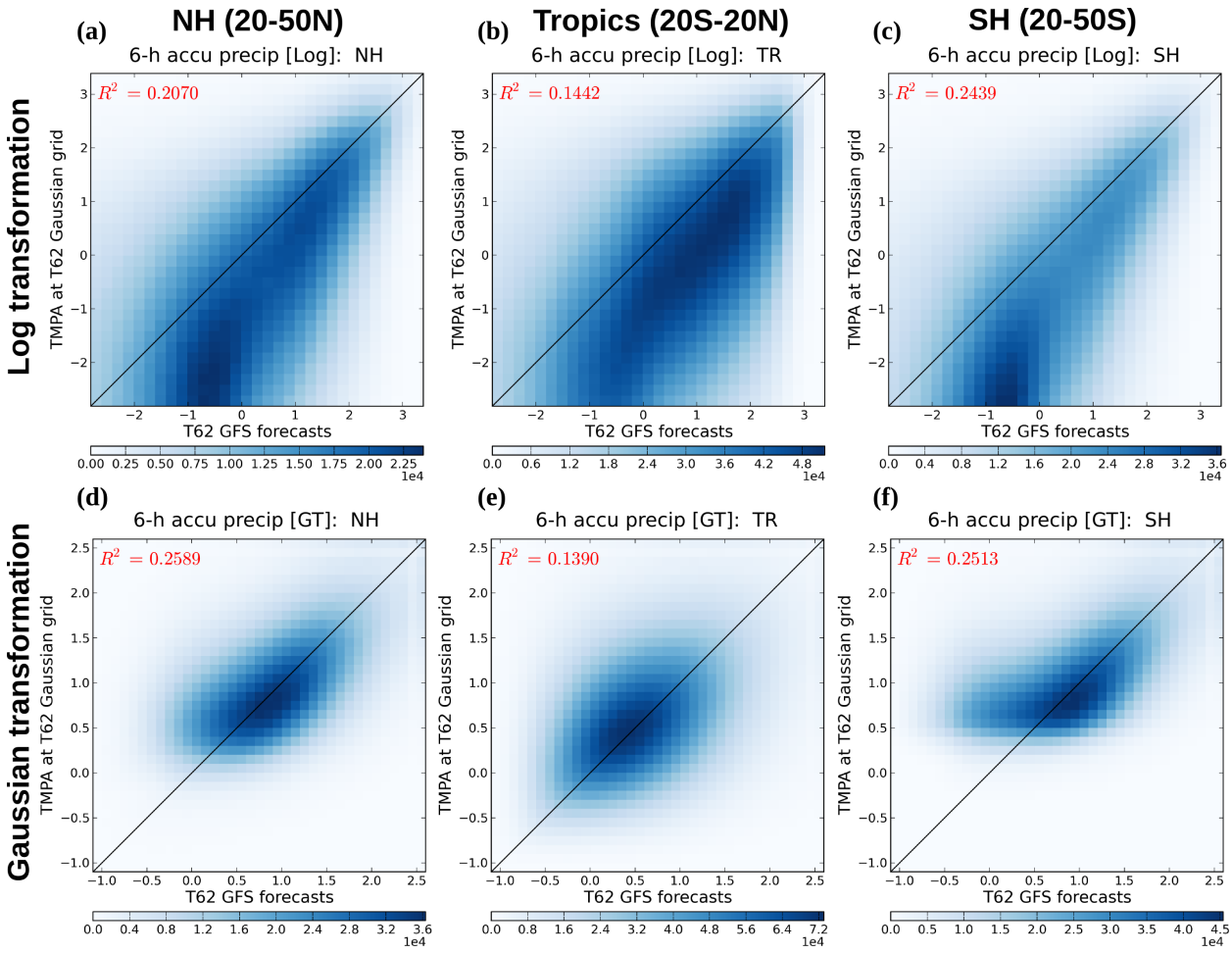

Figure 9: As Figure 8, but for (a) (d) the Northern Hemisphere extratropics (20–50°N), (b) (e) the tropical regions (20°N–20°S), and (c) (f) the Southern Hemisphere extratropics (20–50°S).

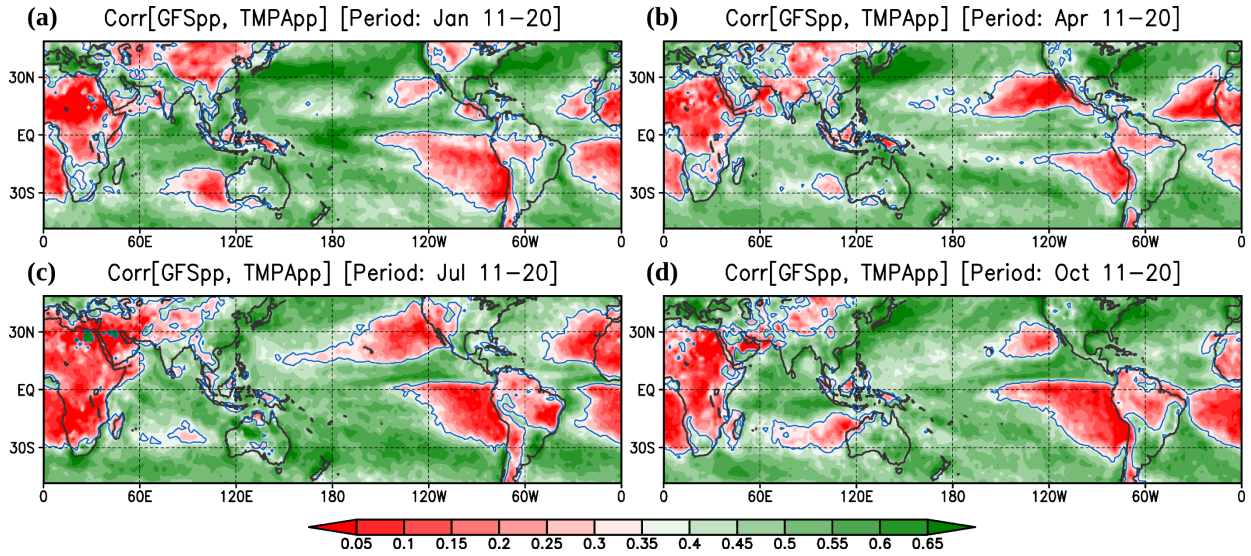

Figure 10: The maps of correlation between precipitation in the GFS model background and in the TMPA observations during the periods of (a) 11–20 January, (b) 11–20 April, (c) 11–20 July, and (d) 11–20 October. The blue contours indicate correlations = 0.35, which is the threshold used for the precipitation assimilation in LMK2015b.